Too Good to be False: Non-Significant Results Revisited

Chris H.J. Hartgerink1

1 Tilburg University, the Netherlands

WORD COUNT: XXXX

Author note

This paper is version controlled and all research files are publicly available at <https://osf.io/qpfnw/>. Analysis code was pre-registered.

Abstract

\*placeholder\*

*Keywords:* p-value, nhst, underpowered, effect size, fisher

**Too Good to be False: Non-Significant Results Revisited**

*Scientific* hypothesis testing has become a well-accepted practice in the psychological sciences, as Popper's (1959/2005) falsificationism has become deeply rooted in current day philosophy of science. Falsification serves as one of the main demarcating criteria of science, although it has its connotations (Maxwell, 1972). Nonetheless, scientific hypothesis testing relies heavily on the falsification property and *statistical* hypothesis testing is borne out of *scientific* hypothesis tests.

In quantitative research, *statistical* hypothesis tests are ubiquitous and required to implement *scientific* hypothesis testing, but these are not equal. The most common *statistical* hypothesis tests are Null Hypothesis Significance Tests (NHST), where a point hypothesis[[1]](#footnote-2) is tested against an alternative hypothesis. The probabilistic nature of such *statistical* hypothesis testing yields *individual* results that are (1) inconclusive, (2) incomplete, and (3) error-prone (i.e., false positives [Type I], and false negatives [Type II]), which are nuances that do not invalidate the paradigm, but require attention when interpreting results (for a review see Harlow, Mulaik, & Steiger, 1997). Inconclusive results are seen in structural equation modeling, where multiple models can yield identical results (Kline, 2011). Results are incomplete, as a *P*-value does not state the probability of a hypothesis (Cohen, 1994), and results are error-prone due to sampling fluctuations (see Table 1 for overview).

False positives have been widely discussed in recent years, as events and papers have clearly indicated the plausibility of widespread false positive findings in the literature. Extreme and (hopefully) incidental fraud cases (Hauser, Stapel, Smeesters, Sanna, and Ruggiero) presented undisputed false positive results, but several papers also indicated that widespread questionable research practices (QRPs; John, Loewenstein, & Prelec, 2012) could lead to dramatic increases in false positive rates (Simmons, Nelson, & Simonsohn, 2011). Questionable editorial practices (QEPs; LeBel et al., 2013), publication bias (Francis, 2012a, 2013), systemic low power (Bakker, Van Dijk, & Wicherts, 2012) and excessive degrees of significance (Francis, 2012b, 2014; Schimmack, 2012; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995) were additional flags for concern. Failures to replicate effects, such as the elderly priming effect (original: Bargh, Chen, & Burrows, 1996; replication: Harris, Coburn, Rohrer, & Pashler, 2013), physical and moral cleansing effect (original: Zhong & Liljenquist, 2006; replications: Earp, Everett, Madva, & Hamlin, 2014; Fayard, Bernstein, & Roberts, 2009), and the unlikely ESP effect (original: Bem, 2011; replication: Ritchie, Wiseman, & French, 2012) raised questions on whether these were false positive findings (Yong, 2012). Additionally, questionable editorial practices (QEPs; LeBel et al., 2013) and the need for innovative results (Trafimow, 2013) in journals could inflate the false positive rate in the literature. It should be noted that the field of psychology took up these concerns, is open to changing its practices (Fuchs, Jenny, & Fiedler, 2012), remedies are investigated (e.g., Murayama, Pekrun, & Fiedler, 2013), and change is on its way (e.g., Brandt et al., 2013; Eich, 2013; Journal of Experimental Social Psychology, 2014).

False negatives have been discussed less explicitly, as the problems for false positives were more salient and more pressing. Power (i.e., true positive rate) has been a pressing issue in the discussion nonetheless, as replications were being discussed for detecting possible false positives (e.g., Asendorpf et al., 2013; Brandt et al., 2013). The power for psychological studies has been estimated to be as low as 35% (Bakker et al., 2012), which implies a 65% false negative rate. This is substantive, given that this exceeds all but one of the false positive rates due to QRPs (81.5%, Table 1; Simmons et al., 2011). Possible power problems have been indicated before (e.g., Cohen, 1962), but to no avail (Sedlmeier & Gigerenzer, 1989). Hence, discussing false negatives is important, as these cause possibly valuable research lines to be stopped prematurely (Fiedler, Kutzner, & Krueger, 2012).

This paper focuses on possible false negative findings in non-significant test results, and whether these results actually show the theoretically expected null distribution. Sample results are possibly interpreted as more certain- and error free than they actually are (due to e.g., law of small numbers, Tversky & Kahneman, 1971; overconfidence of statistical power, Bakker, 2014; Greenwald, 1975), which is especially troubling as non-significant results are often interpreted as the null hypothesis being true (Nickerson, 2000). If non-significant results are truly non-effects, they should be expected to show a uniform *P*-value distribution (Murdoch, Tsai, & Adcock, 2008; Sackrowitz & Samuel-Cahn, 1999), and the accompanying effect size distribution. Our research investigates whether reported statistics from 8 flagship journals corroborates such null distributions, or indicates possible false negative results.

**Methods**

**Data summary**

The dataset used was retrieved from the Open Science Framework[[2]](#footnote-3), and includes APA style test statistics extracted from 8 journals. These test statistics were extracted with statcheck (Epskamp & Nuijten, 2013), and originally included a total of XXXX test results (*t*, *r, F, Z,* χ2 and Wald values). As only *t*, *r,* and *F* values allow for direct and comparable effect size computation, these test results were selected (XXXX results; XX% of original). Table 2 summarizes the selected data used for the analyses in this paper. For a more extensive description of the sampling method underlying the dataset, see the Open Science Framework page, linked to in Footnote 2.

**Effect size distribution**

The selected *t*, *F*, and *r* values are readily computed into effect sizes, which form the observed effect distributions. The effect size metric used throughout the analyses is explained variance, in the form of eta-squared (i.e., η2). For the selected *r* values, this only requires taking the square (i.e., *r*2). *F* and *t* values were converted to effect sizes simultaneously, as *t2* values are *F* values. The formula used to compute these effect sizes is



where for squared *t* values, *df1* equals 1, and *df2* equals the original degrees of freedom from the *t*-test. Adjusted effect sizes were computed with the formulae given in the Appendix.

The theoretical null distribution over all test results was simulated by (1) randomly sampling 1,000,000 test results from the dataset with replacement, (2) sampling non-significant *P*-values uniformly between 0 and α (α = .05), (3) and computing the effect size that accompanies the degrees of freedom of the sampled test results. Effect size computation was done equal to computing the observed effect sizes, except no adjusted effect sizes were computed.

To compare the observed distributions with the theoretical null distribution, Kolmogorov-Smirnov tests were used. The Kolmogorov-Smirnov test is a non-parametric goodness-of-fit test for distributions, which is based on the maximum absolute deviation between the independent distributions being compared (Massey Jr., 1951). In this specific case, the fit of the observed effect size distributions (overall, and per journal) with the null effect distribution is of interest. Differences in distributions between journals were not subjected to inferential significance tests, as the data *are* the population of *t*, *r,* and *F* values reported in the journals.

***P*-value distribution**

Uniformity of *P*-values is tested with the Fisher method ().

**Deviation test.** For the purpose of this paper, the deviation test inspects whether non-significant *P*-values deviate from uniformity. As selecting only non-significant *P*-values restricts the range, the selected *P*-values are transformed back into the state space of [0; 1], by



where *pi* is a vector of *P*-values, and α is the selected significance threshold. The resulting *p\**-values are used to calculate the test statistic gamma,



where *k* is the amount of non-significant *P*-values. The resulting test statistic is gamma distributed, with rate parameter 1 and shape parameter *k* (i.e., γ ~ Γ(1,*k*), as the natural logarithm of a uniform variable is an exponentially distributed, and the sum of *k* exponential distributions is gamma distributed.

**Test power simulations.**

**Paper power.**

**Paper level results**

In the current section, an easy to use method is presented to test for presence of an effect across a set of statistical hypothesis tests, requiring only *p-*values. Subsequently, power simulations are presented and descriptive results are given for papers included in the XXXX dataset.

**Fisher’s inexact test.** Fisher’s test () was originally used to test for deviations from uniformity, which is also applicable for *p-*value distributions.As mentioned previously, if the null hypothesis is true, *p*-values are uniformly distributed — this makes the Fisher test an excellent test for inspecting the presence of an effect. However,

**Fisher test power.** To simulate the power of the Fisher test, the original test results in the dataset were bootstrapped. Sixteen hypothetical population effect sizes were imposed, from which non-significant results were simulated for all *t*, *F* and *r* test statistics in the original dataset. Per set of test statistics within a paper, 1000 iterations were run, where each iteration yielded a result for both Fisher tests per paper. Power was computed as the proportion of significant Fisher tests over the 1000 iterations.

For each test statistic, six steps were necessary to simulate the non-significant *p*-value. First, a critical value under the null distribution was needed. Based on the degrees of freedom in the dataset, this was easily calculated. Second, these same degrees of freedom were used to compute the non-centrality parameter, to determine the population distribution based on the imposed effect size. The non-centrality parameter (i.e., δ) was computed as

where

*N* was determined on the basis of the degrees of freedom (i.e., ; ). Third, the area under the curve of the population distribution was determined, where the result would yield a non-significant result (i.e., β). Fourth, a value was uniformly drawn between 0 and the β value that resulted from step three. Fifth, the accompanying test-value was computed, which was, sixth, used to compute the *p­*-value under the null distribution.

Simulation results indicate the power of both tests is highly sufficient, even when only few non-significant results are presented. Figure 2 visually summarizes the results. It is clear the power rapidly increases as a function of effect size

**Descriptive results.** Of the original test results in the dataset, Fisher tests showed substantial indication for underpowered results. Of the XXXX papers in the dataset, XXXX showed significant deviation from uniform *p-*values.

**Applying to a ‘failed’ research line**

References

Footnotes

Table 1

|  |  |  |
| --- | --- | --- |
|  | H0 | H1 |
| ‘H0’ | 1-α  [0.95] | β  [0.20] |
|  | *True negative* | *Type II error* |
| ‘H1’ | α  [0.05] | 1-β  [0.80] |
|  | *Type I error* | *True positive* |

*Note.* Columns indicate the true situation in the population, rows indicate the statistical conclusion based on sample data. The true positive rate is also called power, and the true negative rate is also called XXXX. Values in square brackets are conventionally acceptable values.

Table 2

Table 3

*Note.* Alpha was assumed to be .05 to determine significant or not.

*Figure 1*

Observed effects versus simulated null effects.

*Figure 2*

Visual depiction of power simulations of Fisher tests. Thick line indicates the

1. This point hypothesis is commonly 0 (nil hypothesis), but can be any point value. [↑](#footnote-ref-2)
2. <https://osf.io/dzrtf/> [↑](#footnote-ref-3)